A Speech Quality Classifier based on Tree-CNN Algorithm that Considers Network Degradations

Samuel Terra Vieira, Renata Lopes Rosa, and Demóstenes Zegarra Rodríguez .

Abstract—**Many factors can affect the users' quality of experience (QoE) in speech communication services. The impairment factors appear due to physical phenomena that occur in the transmission channel of wireless and wired networks. The monitoring of users' QoE is important for service providers. In this context, a non-intrusive speech quality classifier based on the Tree Convolutional Neural Network (Tree-CNN) is proposed. The Tree-CNN is an adaptive network structure composed of hierarchical CNNs models, and its main advantage is to decrease the training time that is very relevant on speech quality assessment methods. In the training phase of the proposed classifier model, impaired speech signals caused by wired and wireless network degradation are used as input. Also, in the network scenario, different modulation schemes and channel degradation intensities, such as packet loss rate, signal-to-noise ratio, and maximum Doppler shift frequencies are implemented. Experimental results demonstrated that the proposed model achieves significant reduction of training time, reaching 25% of reduction in relation to another implementation based on DRBM. The accuracy reached by the Tree-CNN model is almost 95% for each quality class. Performance assessment results show that the proposed classifier based on the Tree-CNN overcomes both the current standardized algorithm described in ITU-T Rec. P.563 and the speech quality assessment method called ViSQOL.**

Index Terms—**Speech quality, objective metrics, wireless network, wired network, deep learning, Tree-CNN.**

I. Introduction

Speech mobile applications and the number of mobile devices [1] has been increasing in the last years [2], and according to [3] the mobile device number will increase across the world. Hence, the telecommunication services performance need to be monitored [4] to guarantee a minimum level of users Quality of Experience (QoE) [5]–[9].

In communication channels occur different type of degradations [10], in wired and wireless networks. The Packet Loss Rate (PLR) is a common kind of degradation [11]. PLR values and its model distributions were widely studied [11]–[13] in the context of wired networks. However, in wireless networks, different impairment factors appear due to the transmission channel characteristics [14], such as the diverse obstacles between the transmission and reception points, which originate different phenomenons. The path delays and signal power variations are example of problems at communication in wireless network. These degradation factors originate the fading [15]. Most of research on speech quality focus on wired network parameters, such as packet losses, jitter and delays; however, the wireless channel impairment characteristics and techniques used in wireless communication systems are not related with the speech quality [16], [17].

Speech quality assessment objective methods estimates a Mean Opinion Score (MOS), and they can be classified in three models depending on the input type of the algorithm [18]: (i) Based on speech signal methods, which can be intrusive and non-intrusive methods. Algorithms that use reference and impaired signals are named intrusive method, such as ITU-T Rec. P.862 [19] and P.863 [20]. Algorithms that only use the impaired speech signal is known as non-intrusive method, which the most representative and standardized algorithm is described in the ITU-T Rec. P.563, but it does not work in a proper manner in networks that presents packet losses [13]. In addition, another non-intrusive method was proposed more recently [21], [22], and reached better performance. (ii) Parametric method that uses as inputs parameters related to network, speech codec and acoustic characteristics. (iii) Hybrid method that uses both approaches.

In recent years, several machine learning algorithms [23], such as Deep Learning (DL) algorithms, have been utilized for speech recognition and analysis. Currently, the Recurrent Neural Networks (RNN), Convolutional Neural Network (CNN) and the Restricted Boltzmann Machine (RBM) are very popular methods used in speech [24] and image recognition reaching satisfactory performance results. RBM is a generative stochastic Artificial Neural Network (ANN) that can learn a probability distribution; for classifications purposes, it is necessary to add a supervised learning method, classifying the samples based on the characteristics extracted by the RBM. Studies regarding the characteristics identification in speech signals demonstrate superior accuracy of the RBM in relation to other widely used methods [25], [26], such as Support Vector Machine (SVM). In a previous work [26], a non-intrusive speech quality classifier based on Discriminative Restricted Boltzmann Machines (DRBM) is presented, and it reached a high accuracy for classifying a MOS speech quality. However, the DRBM presents some deficiencies in terms of the training time, and some proposals try to decrease the impact of this

Manuscript received March 15, 2020; revised April 28, 2020. Date of publication June 4, 2020. Date of current version June 4, 2020. The associate editor prof. Mladen Russo has been coordinating the review of this manuscript and approved it for publication.

This work was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) under Grant 2015/24496-0 and Grant 2018/26455-8.

The part of this paper was presented at the International Conference on Software, Telecommunications and Computer Networks (SoftCOM) 2019.

S. T. Vieira, R. L. Rosa and D. Z. Rodríguez are with the Department of Computer Science, Federal University of Lavras, MG, Brazil, e-mails: [samuel, renata.rosa, demostenes.zegarra]@ufla.br

Digital Object Identifier (DOI): 10.24138/jcomss.v16i2.1032

problem [27]–[29]. More recently, a methodology called Tree-CNN appears to minimize the training time, because in its algorithm, the ANN grows as a tree manner to classify new classes of data, but maintaining the ability to distinguish the previously trained classes at the same time. [30].

It is important to note that, currently, a considerable number of studies [31]–[34] uses machine learning algorithms for speech recognition. However, there is a lack of research on treating the speech quality in communication systems with a high accuracy and which works with significant reduction of training effort.

In this context, this research presents additional contributions regarding to our previous work [26], which can be stated as follows:

- To propose a speech quality assessment method based on the Tree-CNN algorithm for classifying a MOS index into a predefined speech quality class, in a network scenario containing wired and wireless transmission channels. Nowadays, the Tree-CNN has been applying only for images classification with no study cases on speech quality classification.
- Accuracy evaluation of the Tree-CNN in relation to other algorithms, such as, SVM, DRBM and HDRBM.
- Study the complexity of the Tree-CNN in relation to training issues time. The significance of the use of Tree-CNN algorithm is its high accuracy, but mainly the reduced time in the learning process
- Performance assessment of the proposed classification model in relation to ITU-T Rec P.563 and solutions proposed in [21], [26], [35].

To determine the proposed model, different speech signal characteristics are used. In this work, different speech impairment originated by wired and wireless networks are considered. In order to evaluate the impact of those degradation, a test scenario was implemented, in which different packet losses patterns were implemented. The impaired speech sequences were evaluated using the algorithms described in the recommendations ITU-T P.863 [20] and P.563 [35]. The first one is used as reference of speech quality and was an input in the training phase. The later was used for comparing the performance of the proposed speech classifier model based on the Tree-CNN method.

In this research, the Wideband Adaptive Multi-Rate (AMR-WB) [36] and the Enhanced Voice Service (EVS) [37] codecs are used as speech compression algorithms in the implemented test scenario. The AMR-WB is a wideband speech audio coding standard developed based on Adaptive Multi-Rate encoding based on the algebraic code excited linear prediction (ACELP) algorithm. The EVS is the first 3GPP communication codec providing both super-wideband (SWB) and fullband (FB) to improve speech perceptual quality. The AMR-WB e EVS codecs were implemented because they are the most used in current communication networks, specifically, AMR-WB codec is widely used in 3G and 4G networks [38]–[41], and EVS codec [42] is being implemented in the first 5G networks.

The remainder of this article is structured as follows. Section II presents a review of speech characterization and classification models. In section III the impact of wireless channels on the speech quality is presented. Section IV presents proposed classifier model. Section V presents the experimental results. Finally, the conclusions are presented in section VI.

II. Speech Characterization and Classification Models

Speech signals present characteristics that are usually represented by different features [43]; these characteristics extracted from the signal in both temporal and frequency domains.

Nowadays, there are several speech signal features used for different applications. One feature is the Zero Crossing Rate parameter (ZCR), which indicates the speech signal changes during a period of time, from positive to negative amplitude values or vice-versa. The Mel-frequency cepstral coefficients (MFCC) are other features, used by several studies to represent the speech signal [44], [45]. The perceptual linear prediction (PLP) coefficients are utilized for representing the speech spectrum by a compact set of linear prediction coefficients using the Bark frequency scale [46]. In addition to these features, other feature, such as line spectral frequency (LSF) [47], Line Spectral Pairs (LSPs) [48], the spectral centroid, spectral shift, spectral flux, FFT Spectrum information are used for recognition and classification of speech signals.

Unknown speech patterns can be found by unsupervised learning approach, it has been used in several applications [49]–[51]. The RBM can be used as an unsupervised learning approach, which is composed by visible and hidden units; it can learn several discrimination characteristics for a particular problem [52], and eventually improve the computational cost and time required to complete the training process. The main idea of the RBM is to feed the network with unclassified examples and then rebuild the input data. The work of [53] highlights the use of Contrastive Divergence (CD) as a commonly used method for learning in RBMs, because of its efficiency and reliable results. The CD intends to adjust the input values in the model, working the approximation of the learning of maximum likelihood.

The RBM can model fragments of a signal [54] and the RBM can also be used for supervised techniques [26]. The supervised learning was proposed by the DRBM algorithm, in which labels or classes information are incorporate into the visible layer (input); thus, the joint distribution of the input data are calculated and they are classified in a corresponding label. However, the trainning phase depending of the scenario can be more complex that using more simple machine learning algorithms.

A. Tree-CNN

Initial layers of a CNN learn generic features [55] and this characteristic has been used for transferring learning data [56]. In a hierarchical CNN, the upper nodes commonly classify the classes using basic features [57], decreasing the complexity in the training phase.

The Tree-CNN starts as a single root node and after new hierarchies are generated to accommodate new classes. A similar topology is applied in [58], where the new classes are added to the old classes, divided into two super-classes using an error-based model. In [30] is also used the Tree-CNN, however the topology is applied in a totally different scenario, testing images of cars and the results of accuracy are superior to 85%. In this research, our main contribution is to adapt and test the Tree-CNN topology in a scenario of speech perceptual quality classification.

In general, the steps for the Tree-CNN are described as follows:

- Initially, the network is trained for classifying the data into N classes. The data belonging to a new class is presented to the network, then the network grows to accommodate the new class.
- The network grows by adding a new leaf/branch node to the current structure.
- The objective for reducing the training effort is made up of two components, the number of weights updated, and the number of examples, old or new one, required for training.
- Finally, the updates are localized to a new section of the tree.

In the Tree-CNN, the $P(x_t, Tr)$ represents the probability of the input data being classified into the correct category by the trunk-net. There are two sub-networks, one for encoding the input function at a fixed number of class, which is the leaf or branch-net, and another for encoding the locations for the output functions, the trunk-net. The function $P(C_i, Tr)$ represents the probability of the input data being classified into the correct i-esim category. The $P(x_t, Br)$ is the probability of the input data being classified into the i-esim correct category by branch-net. In which, $P(x_t, Br) > P(x_t, Tr)$, because the branch-net is responsible to distinguish the i-esim category from other similar categories, which is different from the Trunk-Net. Assuming that the $P(C_i, Tr)$ is almost iqual to 1, then the probability of the Tree-CNN $P(x_t, Br)P(C_i, Tr)$ classifying the class into the correct category is greater than the probability of the original net $P(x_t, Tr)$. Fig. 1 presents the root and nodes, in which the branch nodes are the intermediaries, having a parent and two or more children and the leaf node represents the last level of the tree; the samples represents the speech files that are classified by the Tree-CNN model.

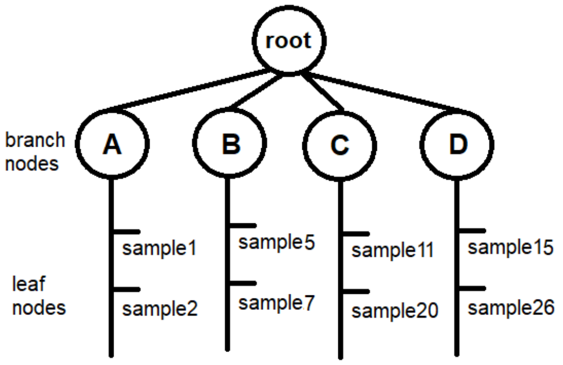


Fig. 1. Tree-CNN topology for classifying the speech samples according to their audio perceptual quality.

III. Wireless Channel Characteristics on Speech Quality

For evaluating the effect of wireless channel degradations on speech quality a test scenario was implemented. The scenario follows the steps:

- Initially, there is an original speech signal (.wav), which is coded by the AMR-WB or EVS codecs;
- Different PLR distributions are applied in the speech signal;
- A modulation scheme is applied in the transmission system that have an impact in the speech signal;
- Modulated signal is transmitted via a RF channel model;
- A demodulation scheme occurs and the speech signal is demodulated and then decode by the AMR-WB or EVS codecs;
- An impaired speech signal is obtained, and the speech quality assessment occurs obtaining a MOS index using both objective algorithms described in ITU-T recommendations P.863 and P.563.

The speech samples extracted from Annex C of ITU-T Recommendation P.501 [59] are used in the tests, which are FB and they have a sampling frequency rate of 48 kHz. Each file has a duration of 8 s, with an initial silent of 0.5 s and intermediate silence of 1 s between two speech segments. It is important to note that he amount of speech activity is greater than 3.2 s in accordance with [60]. The wide-band speech samples are used, to this end an appropriated filter was used. Then, the speech level was equalized to 26 dBov using [61], and a down-sampling process was applied to get frequency sample rate of 16 kHz.

The speech signal passes over a wired network, with different PLR values. The modulation schemes used were the BPSK, QPSK and QAM (QAM-16, QAM-64 and QAM-256). In the wireless transmission channel, the Rayleigh fading channel model was used, in which the parameters configured were SNR (dB) and the maximum Doppler frequency shift (Hz). Table I shows the parameters and respective values used in the simulation.

TABLE I
Simulation Parameter Values

Parameter	Values
PLR (%)	0, 2, 4, 7, 10, 15, 20
SNR (dB)	0:1:30
Max. Doppler Shift (Hz)	0, 5, 20, 50, 100, 200
Modulation scheme	BPSK, QPSK and QAM-(16, 64, 256)

In the reception point, speech samples with different impairments are obtained for comparing with the original speech samples. The ITU-T Rec. P.863 was used to determine the speech quality because the P.863 considers many features related to modern communication systems.In the total $6,510$ different network scenarios are obtained for testing with the AMR-WB codec and the same quantity for the EVS codec. Each simulation was performed 50 times. The data set was separated into 3 sets: training, validation, and testing. 60% of samples corresponding to each network scenario were randomly separated for the training phase, 20% were separated for validation and 20% were used for testing.

At the end, the tests showed that the higher was the SNR, the higher was the MOS index. In case of the maximum Doppler shift, its values do not have a significant effect on the MOS index being very little affected.

IV. Proposed Speech Quality Assessment

Fig. 2 introduces the high level of the network architecture used in this work, using the proposed speech quality classifier based on Tree-CNN. A database was built with the first speech samples database (60%), which was used in the training phase. Different speech characteristic features were extracted from signals and the information is used by the Tree-CNN for determining the Speech Quality Classifier model. The model is equivalent to a subjective rating, which is classified automatically by a machine learning model. The characteristics extracted and used in this work were: ZCR, FFT Spectrum, MFCCs with 13 static characteristics of the MFCC, and the first and second order derivatives of static characteristics, and spectral centroid, spectral flux and spectral shift.

The server is responsible for receiving telephone calls periodically from the service provider to update the database with new speech samples with different types of degradations. After, the learning model is retrained and it performs the extraction of parameters for obtaining new speech characteristics for improving the model. After, the new model is determined, it is sent, by an external application, to the client device, where can be used as a nonintrusive speech quality metric. The users devices are represented by the variables U_1 and U_2; in which the speech signal is analyzed using the updated model for determining the quality class A, B, C or D.

As can be observed in Fig. 2, there are four speech quality classes, which are being considered in this work. They are based on the Absolute Category Rating (ACR) of 5-point scale described in ITU-T Rec. P.800 [62]. At this point, it is important to note that the minimum score, given by P.863 algorithm, is the MOS of 1.0.

V. Results

This section presents the results and characteristics of the learning model topology of the Tree-CNN used in this work and the performance evaluation of the proposed speech quality classifier.

A. Learning Model Topology of the Tree-CNN

The model used in this work was the hierarchical with multiple CNNs because this topology presented best results than just a single CNN acting as a root node with multiple leaf nodes. A new task is defined as learning to identify the speech quality belonging to new classes. There is a root node of our model with a small sample of speech quality from the new training set as input to this node.

A dimensional matrix is obtained from the output layer with the number of children of the root node. The Softmax likelihood was used in the topology.

As stated before, the training, validation and testing phases of 60%, 20% and 20% were considered, respectively. The extraction of sixty-three characteristics of the speech signal in frames was performed in this work. These characteristics were labeled with the classifier after passing through the Tree-CNN, which generated the estimated value for each of the degraded speech samples.

For comparison purposes, other algorithms, such as the SVM, DRBM and HDRBM were also implemented, in order to measure their accuracy. After initial experiments, the Tree-CNN topology with better results was determined, the main parameters of that topology was a linear learning rate of 0.0004, a decay factor of 0.001, and momentum of 0.9. Also, each network of the algorithm was trained using 100 epochs.

B. Performance Evaluation of the Proposed Speech Quality Classifier

The results showed that the Tree-CNN presented results almost equal to the DRBM and HDRBM algorithms, reaching better results, but no a significant improvement. However, using the Tree-CNN was achieved a significant reduction of training effort, which represent a reduction of 25% compared with the DRBM that was used in our previous work [26]. This reduction is very relevant for our proposed solution that need to learn the changes in the network conditions in a fast manner.

Fig. 3 presents the accuracy values in percentage of the classifier algorithms for each speech quality class used in this work for the AMR-WB codec.

Fig. 4 presents the accuracy values in percentage of the classifier algorithms used in this work for the EVS codec.

As expected, results presented in Fig. 3 and Fig. 4 are similar, because the inserted channel transmission degradation in the physical layer were using the same network parameters, and also both codecs have a good response to network fails. In addition, note that Tree-CNN obtained a better performance results that DRBM that was used in our previous work [26].

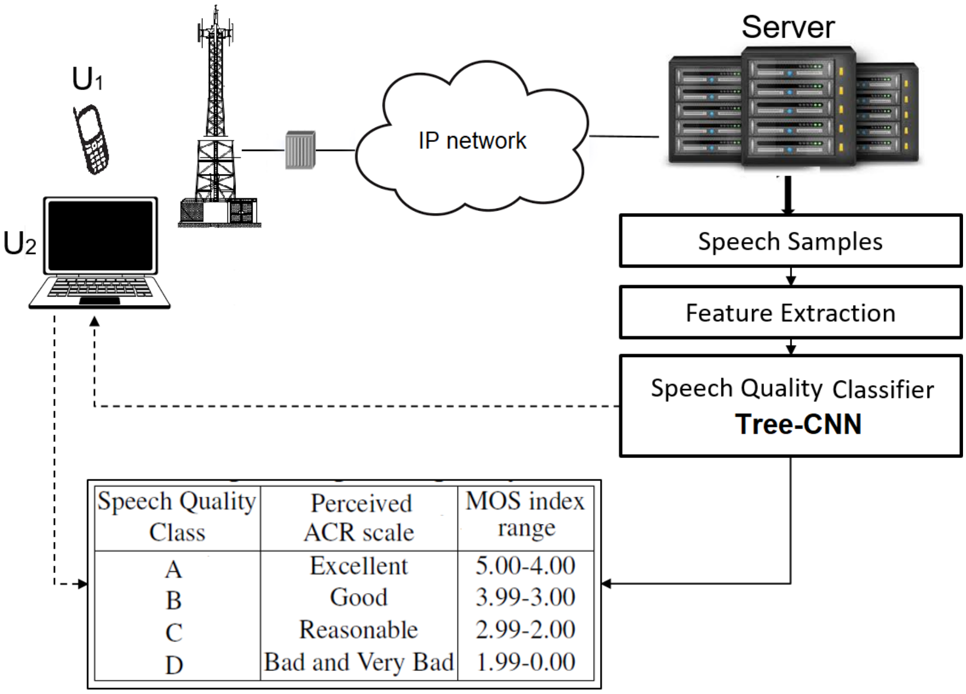
As previously stated, our proposed solution is based on Tree-CNN algorithm, and its classification accuracy is compared with two non-intrusive methods, one of them is ITU-T Rec. P.563 and the other one is the solution descried in [21].

The performance evaluation results of the proposed model correspond to the validation phase.

Firstly, the results are presented considering the accuracy of the proposed classifier and the ITU-T Rec P.563.

Confusion Matrix is a performance measurement for machine learning classification, and it was used to present the results in this work. The class determined by the proposed model is performed automatically, by the Tree-CNN, and the P.563 MOS scores are adequate to the corresponding speech quality. Table II presents the confusion Matrix results of both algorithms, the P.563 and the classifier model, named of TCNN, for the tests using the AMR-WB codec.

Similarly, Table III presents the accuracy on speech quality predictions of proposed classifier model and the P.563 algorithm for the tests using the EVS codec. The results show that the classifier model has a better prediction than the P.563 algorithm.



Speech Quality Class	Perceived ACR scale	MOS index range
A	Excellent	5.00-4.00
B	Good	3.99-3.00
C	Reasonable	2.99-2.00
D	Bad and Very Bad	1.99-0.00

Fig. 2. Network architecture of the proposed solution regarding the speech quality classification method.

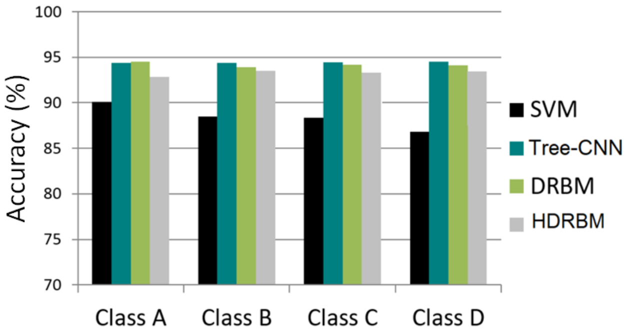


Fig. 3. Accuracy of SVM, DRBM and HDRBM and Tree-CNN for the tests using the AMR-WB codec.

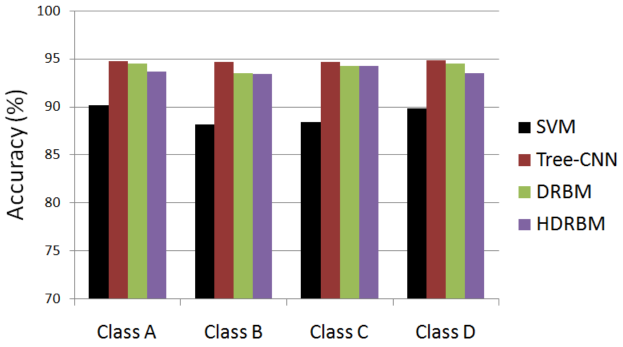


Fig. 4. Accuracy of SVM, DRBM and HDRBM and Tree-CNN for the tests using the EVS codec.

TABLE II
SPEECH QUALITY CLASS PREDICTION (%) USING THE TREE-CNN (TCNN) MODEL AND THE P.563 ALGORITHM FOR THE AMR-WB

Speech Qual. Class	TCNN / P.563 A	TCNN / P.563 B	TCNN / P.563 C	TCNN / P.563 D
A	94.21 / 51.98	5.79 / 16.52	0.0 / 29.44	0.0 / 2.06
B	3.52 / 1.25	94.89 / 55.27	1.59 / 38.74	0.0 / 4.74
C	0.0 / 0.58	1.55 / 6.51	95.01 / 88.24	3.44 / 4.67
D	0.0 / 0.0	0.0 / 3.25	6.11 / 7.22	93.89 / 89.53

TABLE III
SPEECH QUALITY CLASS PREDICTION (%) USING THE TREE-CNN MODEL AND THE P.563 ALGORITHM FOR THE EVS

Speech Qual. Class	TCNN / P.563 A	TCNN / P.563 B	TCNN / P.563 C	TCNN / P.563 D
A	92.11 / 50.33	7.89 / 15.11	0.0 / 27.89	0.0 / 6.67
B	2.33 / 1.48	95.99 / 54.37	1.68 / 39.13	0.0 / 5.02
C	0.0 / 0.63	2.17 / 6.28	95.11 / 87.33	2.72 / 5.76
D	0.0 / 0.0	0.0 / 5.72	6.22 / 8.19	93.78 / 86.09

It can be observed from Table II and Table III that the proposed Tree-CNN model largely overcomes the ITU-T P.563. It is important to note that P.563 algorithm was not tested in wireless context and it is recommended for NB networks.

In addition, in order to compare the proposed model based on Tree-CNN with another speech quality metric, the solution introduced in [21] was used because it is available and get reliable results. There are other solutions that were not considered because they are parametric models or algorithms are not available [63]–[67] .

Table IV presents the results reached by our model and the solution presented in [21], in the scenarios in which the AMR-WB codec was used.

Table V presents the results reached by our model and the solution presented in [21], in the scenarios in which the EVS

TABLE IV
SPEECH QUALITY CLASS PREDICTION (%) USING THE TREE-CNN (TCNN) MODEL AND THE VISQOL [21] ALGORITHM FOR THE AMR-WB

Speech Qual. Class	TCNN / VisQOL A	TCNN / VisQOL B	TCNN / VisQOL C	TCNN / VisQOL D
A	94.21 / 89.85	5.79 / 8.45	0.0 / 1.16	0.0 / 0.54
B	3.52 / 6.52	94.89 / 90.15	1.59 / 3.33	0.0 / 0.0
C	0.0 / 1.2	1.55 / 3.16	95.01 / 86.12	3.44 / 9.52
D	0.0 / 0.4	0.0 / 2.33	6.11 / 10.25	93.89 / 87.12

codec was used.

TABLE V
SPEECH QUALITY CLASS PREDICTION (%) USING THE TREE-CNN MODEL AND THE VISQOL [21] ALGORITHM FOR THE EVS

Speech Qual. Class	TCNN / VisQOL A	TCNN / VisQOL B	TCNN / VisQOL C	TCNN / VisQOL D
A	92.11 / 88.22	7.89 / 9.26	0.0 / 2.52	0.0 / 0.0
B	2.33 / 7.26	95.99 / 89.10	1.68 / 3.64	0.0 / 0.0
C	0.0 / 0.80	2.17 / 2.80	95.11 / 88.25	2.72 / 8.15
D	0.0 / 0.0	0.0 / 3.10	6.22 / 9.78	93.78 / 87.12

From Table IV and Table V, it is worth to note that [21] was not trained with the same network impairments that our proposed model, and this fact could have influenced in its performance results.

Finally, subjective tests were performed in a controlled environment with 31 volunteers, in which 12 were women and 19 were men, aged between 17 and 43 years. Each assessor reported having no experience in speech quality testing in general. The evaluations time were carried out during 5 weeks.

Each volunteer analyzed at least 20 sample files using a 5-point quality scale. The experimental results showed that the proposal model using the Tree-CNN reached an accuracy of 93.8 %.

VI. CONCLUSION

In wireless scenarios, is common that many types of degradations occur in the communication system, as a consequence, the speech quality is affected and it must be measured. We used a database of impaired speech samples caused by Doppler shift, SNR, and PLR, trying to cover wired and wireless network degradation. In this work, a simulator was built, considering, the AMR-WB and EVS speech codec, using different modulation schemes, and different wired and wireless channel impairment conditions. The results present the high relation between the network degradation and the speech perceptual quality. Different machine learning algorithms were tested, including the DRBM algorithm that was used in a previous work [26], and the Tree-CNN algorithm obtained the highest accuracy results. Most relevant, experimental results demonstrated that Tree-CNN achieved significant reduction of training time in all scenarios including both AMR-WB and EVS codec that is very important for speech quality metrics. Based on these results the proposed speech quality classifier is built using the Tree-CNN algorithm for classifying speech quality samples. In the performance validation tests, results showed that our proposed classifier model with the Tree-CNN was more efficient than the P.563 and VisQOL algorithms, in all tested scenarios. Furthermore, the validation results obtained by subjective tests indicated that the proposed model reached a classification accuracy of 93.80%.

REFERENCES

[1] H. Fall, O. Zytoune, and M. Yahyai, “Theory of algorithm suitability on managing radio resources in next generation mobile networks,” *Journal of Communications Software and Systems*, vol. 14, no. 2, pp. 180–188, 2018.

[2] M. Christian and et al., “Smartphone usage in the 21st century: who is active on whatsapp?,” *BMC Research Notes*, vol. 8, no. 1, pp. 331–336, Aug. 2015.

[3] Cisco Inc., “Visual networking index: Global mobile data traffic forecast update, 2016–2021,” June 2017.

[4] D. Rodriguez, G. Pivaro, and J. Sousa, “Apparatus and method for evaluating voice quality in a mobile network,” in *Patent number US 9,078,143 B2 by US Patent and Trademark Office*, Jul. 2015, pp. 1–15.

[5] D. U. U. Garip, O. Çalık, and G. K. Kurt, “Impact of retransmissions on the quality of experience with realistic channel model,” in *2017 International Symposium on Wireless Communication Systems (ISWCS)*, Aug 2017, pp. 299–302.

[6] D. Z. Rodríguez, R. L. Rosa, E. A. Costa, J. Abrahão, and G. Bressan, “Video quality assessment in video streaming services considering user preference for video content,” *IEEE Transactions on Consumer Electronics*, vol. 60, no. 3, pp. 436–444, Aug 2014.

[7] E. L. Lasmar, F. O. de Paula, R. L. Rosa, J. I. Abrahão, and D. Z. Rodríguez, “Rsrs: Ridesharing recommendation system based on social networks to improve the user’s qoe,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 12, pp. 4728–4740, 2019.

[8] R. L. Rosa, D. Z. Rodriguez, and G. Bressan, “Sentimeter-br: A social web analysis tool to discover consumers’ sentiment,” in *2013 IEEE 14th International Conference on Mobile Data Management*, June 2013, vol. 2, pp. 122–124.

[9] Ines Ramadža, Vesna Pekić, and Julije Ožegović, “Quality of experience driven network performance criteria for telecommunication traffic types,” *Journal of Communications Software and Systems*, vol. 15, 07 2019.

[10] Shun Kojima, Kazuki Maruta, and Chang-Jun Ahn, “Throughput maximization by adaptive switching with modulation coding scheme and frequency symbol spreading,” *Journal of Communications Software and Systems*, vol. 14, no. 4, pp. 332–339, 2018.

[11] E. T. Affonso, R. L. Rosa, and D. Z. Rodríguez, “Speech quality assessment over lossy transmission channels using deep belief networks,” *IEEE Signal Processing Letters*, vol. 25, no. 1, pp. 70–74, Jan 2018.

[12] J. Saldana, J. Fernández-Navajas, J. Ruiz-Mas, E. Viruete Navarro, and L. Casadesus, “The utility of characterizing packet loss as a function of packet size in commercial routers,” in *2012 IEEE Consumer Communications and Networking Conference (CCNC)*, Jan 2012, pp. 346–347.

[13] J. Polacky and P. Pocta, “An analysis of the impact of packet loss, codecs and type of voice on internal parameters of P.563 model,” in *Proc. Int. Conf. on Digital Technlogies*, Slovakia, July 2014, pp. 281–284.

[14] Pedro Sousa, Vitor Pereira, Paulo Cortez, Miguel Rio, and Miguel Rocha, “A framework for improving routing configurations using multi-objective optimization mechanisms,” *Journal of Communications Software and Systems*, vol. 12, no. 3, pp. 145–156, 2017.

[15] E. T. Affonso, D. Z. Rodríguez, R. L. Rosa, T. Andrade, and G. Bressan, “Voice quality assessment in mobile devices considering different fading models,” in *Proc. IEEE International Symposium on Consumer Electronics (ISCE’2016)*, Sept. 2016, pp. 21–22.

[16] D. Z. Rodríguez, G. F. Pivaro, R. L. Rosa, G. Mittag, and S. Möller, “Quantifying the quality improvement of mimo transmission systems in voip communication,” in *Proc. 26th International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, Sept 2018, pp. 1–5.

[17] D. Z. Rodríguez, G. F. Pivaro, R. L. Rosa, G. Mittag, and S. Möller, “Improving a parametric model for speech quality assessment in wireless communication systems,” in *Proc. 26th International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, Sept 2018, pp. 1–5.

[18] S. Möller, W. Y. Chan, N. Côté, T. H. Falk, A. Raake, and M. Wältermann, “Speech quality estimation: Models and trends,” *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 18–28, Nov 2011.

[19] ITU-T Rec. P.862, “Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs,” Feb. 2001.

[20] ITU-T Rec. P.863, "Perceptual objective listening quality assessment (POLQA)," Sep. 2014.

[21] C. Sloan, N. Harte, D. Kelly, A. C. Kokaram, and A. Hines, "Objective assessment of perceptual audio quality using ViSQOLAudio," *IEEE Transactions on Broadcasting*, vol. 63, no. 4, pp. 693–705, Dec 2017.

[22] Andrew Hines, Jan Skoglund, Anil C. Kokaram, and Naomi Harte, "Robustness of speech quality metrics to background noise and network degradations: Comparing ViSQOL, PESQ and POLQA," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'2013)*, Vancouver, BC, Canada, May 2013, pp. 3697–3701.

[23] Z. Fan and R. Liu, "Investigation of machine learning based network traffic classification," in *2017 International Symposium on Wireless Communication Systems (ISWCS)*, Aug 2017, pp. 1–6.

[24] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 6645–6649.

[25] B. Yoshua, C. Nicolas, D. Olivier, L. Hugo, S-M. Xavier, H. Christian, and L. Jérôme, "Detonation classification from acoustic signature with the restricted boltzmann machine," *Computational Intelligence*, vol. 28, no. 2, pp. 261–288, 2012.

[26] D. Militani, D. C. Begazo, R. Rosa, and D. Z. Rodríguez, "A speech quality classifier based on signal information that considers wired and wireless degradations," in *2019 Int. Conf. on Software, Telecommunications and Computer Networks (SoftCOM)*, Sep. 2019, pp. 1–6.

[27] C. Plahl, T. N. Sainath, B. Ramabhadran, and D. Nahamoo, "Improved pre-training of deep belief networks using sparse encoding symmetric machines," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4165–4168.

[28] C. Sloan, N. Harte, D. Kelly, A. C. Kokaram, and A. Hines, "Stochastically reducing overfitting in deep neural network using dropout," *International Journal of Innovative Science, Engineering Technology*, vol. 2, no. 5, pp. 465–469, May 2015.

[29] L. Chen, L. Yang, C. Sun, and H. Xi, "A fast rbm-hidden-nodes based extreme learning machine," in *2017 29th Chinese Control And Decision Conference (CCDC)*, 2017, pp. 2121–2126.

[30] Deboleena Roy, Priyadarshini Panda, and Kaushik Roy, "Tree-cnn: A hierarchical deep convolutional neural network for incremental learning," *Neural Networks*, vol. 121, pp. 148 – 160, 2020.

[31] G. Zweig, C. Yu, J. Droppo, and A. Stolcke, "Advances in all-neural speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 4805–4809.

[32] H. Erdogan, J. Hershey, S. Watanabe, and J. Le Roux, *Deep Recurrent Networks for Separation and Recognition of Single-Channel Speech in Nonstationary Background Audio*, pp. 165–186, Springer International Publishing, Cham, 2017.

[33] F. L. de Almeida, R. L. Rosa, and D. Z. Rodriguez, "Voice quality assessment in communication services using deep learning," in *2018 15th International Symposium on Wireless Communication Systems (ISWCS)*, Aug 2018, pp. 1–6.

[34] E. T. Affonso, R. D. Nunes, R. L. Rosa, G. F. Pivaro, and D. Z. Rodríguez, "Speech quality assessment in wireless voip communication using deep belief network," *IEEE Access*, vol. 6, pp. 77022–77032, 2018.

[35] ITU-T Rec. P.563, "Single-ended method for objective speech quality assessment in narrow-band telephony applications," May. 2004.

[36] 3GPP TS 26.171, "Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; General Description (v15.0.0)," June 2018.

[37] S. Bruhn, H. Pobloth, M. Schnell, B. Grill, J. Gibbs, L. Miao, K. Järvinen, L. Laaksonen, N. Harada, N. Naka, S. Ragot, S. Proust, T. Sanda, I. Varga, C. Greer, M. Jelínek, M. Xie, and P. Usai, "Standardization of the new 3GPP EVS codec," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 5703–5707.

[38] R. Mullner, C. F. Ball, K. Ivanov, D. Hartmann, and H. Winkler, "Amr-wideband: enjoying superior voice quality at full coverage and competitive capacity in geran networks," in *2005 IEEE 61st Vehicular Technology Conference*, 2005, vol. 4, pp. 2320–2324 Vol. 4.

[39] D. Z. Rodríguez, R. L. Rosa, F. L. Almeida, G. Mittag, and S. Möller, "Speech quality assessment in wireless communications with mimo systems using a parametric model," *IEEE Access*, vol. 7, pp. 35719–35730, 2019.

[40] D. Z. Rodríguez and S. Möller, "Speech quality parametric model that considers wireless network characteristics," in *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, 2019, pp. 1–6.

[41] F. R. Avila and A. N. Barreto, "Non-intrusive speech quality estimation for gsm system using narrowband and wideband amr codec," in *2008 IEEE International Symposium on Wireless Communication Systems*, 2008, pp. 173–177.

[42] D. McGrath, S. Bruhn, H. Purnhagen, M. Eckert, J. Torres, S. Brown, and D. Darcy, "Immersive audio coding for virtual reality using a metadata-assisted extension of the 3gpp evs codec," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 730–734.

[43] S. Nayak, S. Bhati, and K. S. R. Murty, "An investigation into instantaneous frequency estimation methods for improved speech recognition features," in *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Nov 2017, pp. 363–367.

[44] Wei Han, Cheong-Fat Chan, Chiu-Sing Choy, and Kong-Pang Pun, "An efficient mfcc extraction method in speech recognition," in *IEEE International Symposium on Circuits and Systems*, May 2006, pp. 4 pp.–.

[45] Jong-Hwan Lee, Ho-Young Jung, Te-Won Lee, and Soo-Young Lee, "Speech feature extraction using independent component analysis," in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*, June 2000, vol. 3, pp. 1631–1634 vol.3.

[46] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, Apr 1990.

[47] E. Bozkurt, E. Erzin, C. E. Erdem, and A. T. Erdem, "Use of line spectral frequencies for emotion recognition from speech," in *2010 20th International Conference on Pattern Recognition*, Aug 2010, pp. 3708–3711.

[48] K. K. Paliwal, "A study of line spectrum pair frequencies for speech recognition," in *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*, April 1988, pp. 485–488 vol.1.

[49] Y. Koizumi, S. Saito, H. Uematsu, Y. Kawachi, and N. Harada, "Unsupervised detection of anomalous sound based on deep learning and the neyman-pearson lemma," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1–1, 2018.

[50] O. E. Nii Noi, M. Qirong, G. Xu, and Y. Xue, "Coupled unsupervised deep convolutional domain adaptation for speech emotion recognition," in *IEEE Int. Conf. on Multimedia Big Data*, Sept 2018, pp. 1–5.

[51] O.-J. Räsänen, U. Laine, and T. Altosaar, "Self-learning vector quantization for pattern discovery from speech," in *10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 6-10, 2009*, 2009, pp. 852–855.

[52] G. Pan, J. Qiao, W. Chai, and N. Dimopoulos, "An improved RBM based on bayesian regularization," in *Proc. Int. Joint Conf. on Neural Networks*, Beijing, China, Jul. 2014, pp. 2935–2939.

[53] G. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computing*, vol. 18, no. 7, pp. 1527–1554, 2006.

[54] N. Jaitly and G. Hinton, "Learning a better representation of speech soundwaves using restricted boltzmann machines.," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal*, Prague, Czech Republic, May 2011, pp. 5884–5887.

[55] Syed Shakib Sarwar, Priyadarshini Panda, and Kaushik Roy, "Gabor filter assisted energy efficient fast learning convolutional neural networks," *2017 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, Jul 2017.

[56] Syed Shakib Sarwar, Aayush Ankit, and Kaushik Roy, "Incremental learning in deep convolutional neural networks using partial network sharing," 2017.

[57] Zhicheng Yan, Hao Zhang, Robinson Piramuthu, Vignesh Jagadeesh, Dennis DeCoste, Wei Di, and Yizhou Yu, "Hd-cnn: Hierarchical deep convolutional neural network for large scale visual recognition," 2014.

[58] Tianjun Xiao, Jiaxing Zhang, Kuiyuan Yang, Yuxin Peng, and Zheng Zhang, "Error-driven incremental learning in deep convolutional neural network for large-scale image classification," in *Proceedings of the 22nd ACM International Conference on Multimedia*, New York, NY, USA, 2014, p. 177–186, Association for Computing Machinery.

[59] ITU-T Rec. P.501, "Test signals for use in telephonometry," 2017.

[60] ITU-T Rec. P.631.1, "Application guide for recommendation ITU-T P.863," 2014.

[61] ITU-T Rec. P.191, "Software tools for speech and audio coding standardization," Mar. 2010.

[62] ITU-T Rec. P.800, "Methods for subjective determination of transmission quality," Aug. 1996.

[63] R. Dantas Nunes, R. Lopes Rosa, and D. Zegarra Rodríguez, "Performance improvement of a non-intrusive voice quality metric in lossy networks," *IET Communications*, vol. 13, no. 20, pp. 3401–3408, 2019.

[64] S. Zafar, I. F. Nizami, and M. Majid, “Non-intrusive speech quality assessment using natural spectrogram statistics,” in *2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, 2020, pp. 1–4.

[65] D. Z. Rodríguez, M. Arjona Ramírez, L. F. Bernardes, G. Mittag, and S. Möller, “Impact of fec codes on speech communication quality using wb e-model algorithm,” in *2019 Wireless Days (WD)*, 2019, pp. 1–4.

[66] P. Bachhav, M. Todisco, and N. Evans, “Artificial bandwidth extension using conditional variational auto-encoders and adversarial learning,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6924–6928.

[67] Y. Xia, S. Braun, C. K. A. Reddy, H. Dubey, R. Cutler, and I. Tashev, “Weighted speech distortion losses for neural-network-based real-time speech enhancement,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 871–875.

Samuel ST Vieira completed his bachelor’s degree in Computer Science at the Federal Institute of Education, Science and Technology of Minas Gerais, Brazil, 2016. He is currently a master’s student in Computer Science at the Federal University of Lavras with his research in simulating the propagation of radio signals and assessment of voice quality. His research interest includes QoS and QoE in telecommunication services and artificial intelligence algorithms.

Renata L. Rosa received her M.S. degree from the University of São Paulo in 2009 and her Ph.D. degree from the Polytechnic School of the University of São Paulo, in 2015 (EPUSP). She is currently an Adjunct Professor with Department of Computer Science, Federal University of Lavras, Brazil. Her current research interests include computer networks, artificial intelligence algorithms, recommendation systems, telecommunication systems, wireless networks, and quality of service and quality of experience in multimedia services.

Demóstenes Z. Rodríguez (M’12-SM’15) received the B.S. degree in electronic engineering from the Pontifical Catholic University of Peru, the M.S. degree and Ph.D. degree from the University of São Paulo in 2009 and 2013. He is currently an Adjunct Professor with the Department of Computer Science, Federal University of Lavras, Brazil. He has a solid knowledge in Telecommunication Systems and Computer Science based on 15 years of Professional experience in major companies. His research interest includes QoS and QoE in Multimedia services, artificial intelligence algorithms, and architect solutions in Telecommunication Systems.